PSYCHOLOGY

doi: 10.3389/fpsyg.2011.00326

What are the visual features underlying rapid object recognition?

Sébastien M. Crouzet* and Thomas Serre

Cognitive, Linguistic, and Psychological Sciences Department, Institute for Brain Sciences, Brown University, Providence, RI, USA

Edited by:
Rufin VanRullen, Centre de Recherche Cerveau et Cognition, France

Reviewed by:
Hans P Op De Beeck, Catholic University of Leuven, Belgium
Jan Drewes, CerCo, France

****Correspondence:***
Sébastien M. Crouzet, Cognitive, Linguistic, and Psychological Sciences Department, Institute for Brain Sciences, Brown University, Providence, RI 02912, USA.
e-mail: seb.crouzet@gmail.com

Research progress in machine vision has been very significant in recent years. Robust face detection and identification algorithms are already readily available to consumers, and modern computer vision algorithms for generic object recognition are now coping with the richness and complexity of natural visual scenes. Unlike early vision models of object recognition that emphasized the role of figure-ground segmentation and spatial information between parts, recent successful approaches are based on the computation of loose collections of image features without prior segmentation or any explicit encoding of spatial relations. While these models remain simplistic models of visual processing, they suggest that, in principle, bottom-up activation of a loose collection of image features could support the rapid recognition of natural object categories and provide an initial coarse visual representation before more complex visual routines and attentional mechanisms take place. Focusing on biologically plausible computational models of (bottom-up) pre-attentive visual recognition, we review some of the key visual features that have been described in the literature. We discuss the consistency of these feature-based representations with classical theories from visual psychology and test their ability to account for human performance on a rapid object categorization task.

Keywords: rapid visual object recognition, computational models, visual features, computer vision, feedforward

1. INTRODUCTION

Object recognition is concerned with determining the identity of an object in our visual field of view. Such process relies on visual representations that need to be both selective (recognizing our friend among many other faces) and invariant (recognizing our friend irrespective of drastic changes in visual appearance due to changes in position, size, viewpoint, illumination, or even facial expression; Ullman, 1996; Riesenhuber and Poggio, 1999). The computations carried out on these representations feel effortless and almost immediate (our subjective experience suggests that we know what it is that we are looking at as soon as we see it).

Progress in our understanding of the computational mechanisms underlying visual object recognition has been significant, with converging evidence from neuroscience, psychology, and computer science (Serre and Poggio, 2010). Shape and object category information has been traditionally associated with processing in the ventral stream of the visual cortex. A long-standing metaphor for the underlying processes is that of filtering. The princeps discovery was made by Hubel and Wiesel (1959, 1968), who first reported the existence of bar and edge detectors in the primary visual cortices of the mammalian brain. They further proposed the first cortical model of visual processing thereby suggesting that such selectivity for oriented bars could be achieved via selective pooling mechanisms from the spatial arrangements of center-surround ganglion cells in the Lateral Geniculate Nucleus (LGN; Hubel and Wiesel, 1962). These ideas later formed the basis of Marr's *primal sketch* in his prominent computational theory of visual processing (Marr, 1982). Today, edge detection and spatial frequency analysis as the building block of early vision remains the dogma. However, our understanding of subsequent stages of processing along the visual hierarchy remains a matter of debate.

Marr famously postulated that the next stage of visual processing was concerned with the building of intermediate 2(1/2)D representations for surfaces toward the explicit construction of 3D representations for matching stimuli to internal representations of objects and/or storage into memory. These ideas motivated a subsequent theory by Biederman (1987), the recognition-by-components (RBC), which emphasizes the role of figure-ground segmentation and explicit encoding of spatial relations between 3D object parts. These 3D parts, named *geons*, are analogous to syllables in linguistics and constitute a generic vocabulary for representing objects with different combinations and spatial arrangements of these elements. A typical processing pipeline is sketched on **Figure 1** (left) with key stages of visual processing including: edge detection → grouping → segmentation → matching.

Around the same time, several psychophysical studies suggested that a coarse image analysis based on simple feature detectors could be done very rapidly in parallel across the visual field (Treisman and Gelade, 1980; Julesz, 1981; Bergen and Julesz, 1983). The study of what can be seen "at first sight" has since been intensively pursued using the visual search paradigm. Two prominent theories seem to account for most experimental data: the Feature Integration Theory by Treisman and Gelade (1980) and the Guided Search Theory by Wolfe (2006). Both suggest that simple image features such as color, orientation, motion, or size (see Wolfe and Horowitz, 2004 for an extensive review) can be processed pre-attentively and

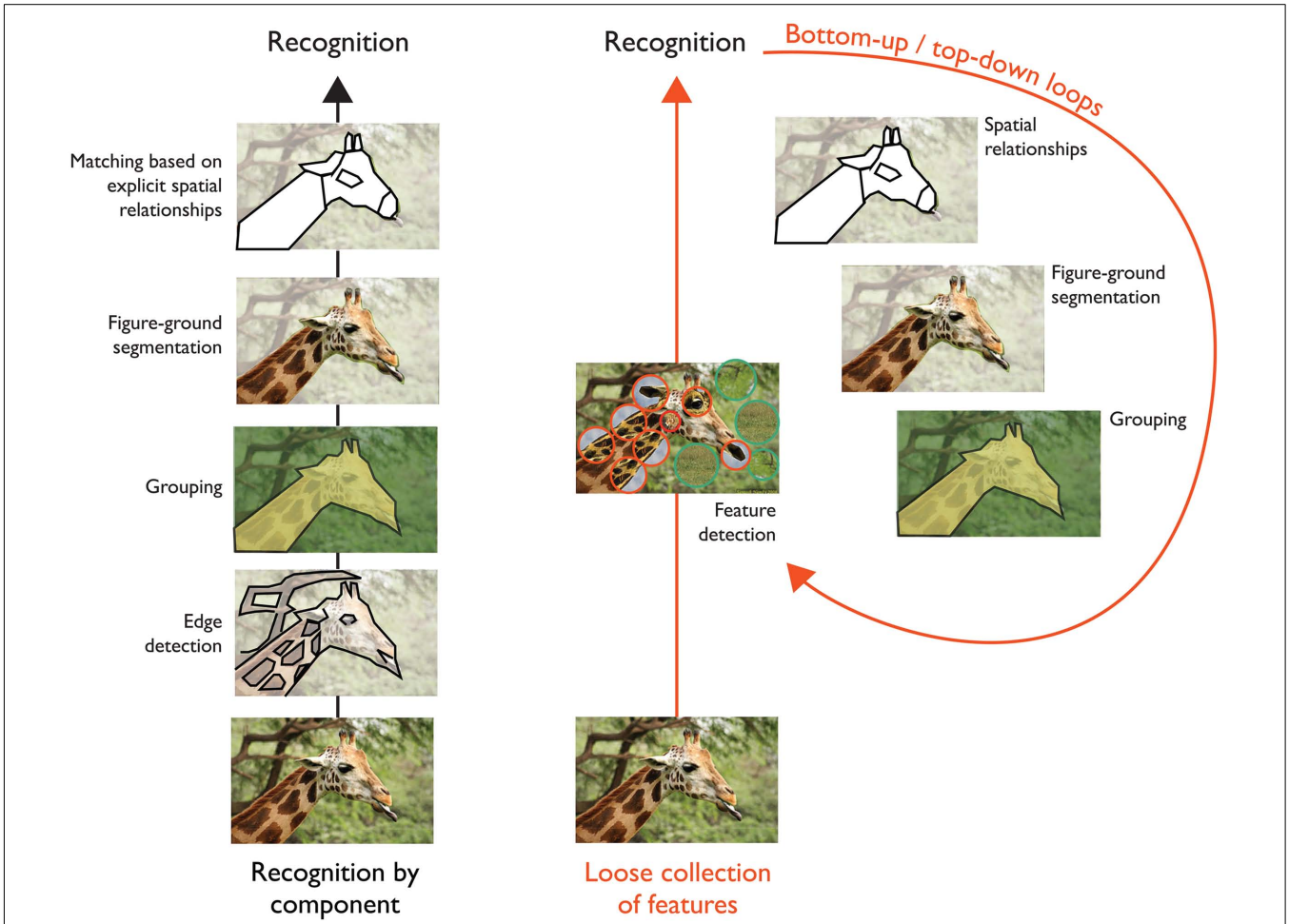

FIGURE 1 | (Left) Overview of the successive steps involved in classical theories of object recognition (see text for details). (Right) An alternative view is that the bottom-up activation of a loose collection of hardwired feature detectors via a hierarchy of increasing complex processing stages may provide a coarse initial visual representation for more complex routines and several feedforward/feedback iterations to solve specific tasks including contours detection, grouping, figure segregation as well as the computation of spatial relations between parts, and more generally, the parsing and interpretation of complex visual scenes (see for instance: Hochstein and Ahissar, 2002; Zheng et al., 2007; Epshtein et al., 2008; Serre and Poggio, 2010 for a review).

in parallel. However, any search for more complex combinations of features (e.g., T among Ls) for which hardwired feature detectors are not readily available will lead to reaction times that are dependent on the number of distractors in the display; a phenomenon consistent with a serial attentional process.

Studies conducted on natural visual scenes came to challenge some of these ideas by demonstrating the incredible speed and accuracy of our visual system for some of the most challenging visual recognition tasks in natural scenes. For instance, the rapid serial visual presentation (RSVP, Potter and Levy, 1969) and the rapid visual categorization (Thorpe et al., 1996) paradigms showed that human subjects are able to recognize (and remember) objects presented very rapidly in the absence of eye movements and potentially, shifts of attention. Further EEG studies measuring event related potentials (ERPs) directly on the scalp showed robust differential activity between target and distractor images within 150 ms after stimulus presentation (VanRullen and Thorpe, 2001). Recent studies using backward-masking (Bacon-Macé et al., 2005) and saccadic responses (Kirchner and Thorpe, 2006; Crouzet et al., 2010) suggest that recognition is possible under even more severe time constraints, possibly via a single feedforward sweep through the visual system (Lamme and Roelfsema, 2000; VanRullen and Koch, 2003). The underlying visual representation remains relatively coarse as it was shown that participants frequently fail to localize targets that they had correctly detected in an RSVP stream (Evans and Treisman, 2005). In particular, this seems inconsistent with recognition processes that rely on explicit encoding of spatial relationships between parts and suggest instead that rapid recognition may rely on the detection of an "unbound" collection of image features.

Consistent with this idea, the rapid recognition of natural object categories such as animals does not seem to require attention: The level of performance of human observers remains high even when two images are flashed simultaneously (Rousselet et al., 2002) and when stimuli are presented in the periphery while an attention-demanding (letter discrimination) task is performed at the fovea

(dual-task paradigm, Li et al., 2002). To account for these results, VanRullen suggested that the recognition of natural object categories must be based on a dictionary of features that are *hardwired* in the visual system (VanRullen, 2007).

This idea is, in fact, consistent with the unsupervised learning mechanisms of feature hierarchies postulated by current models of the visual cortex (see Serre et al., 2007a for a review) and provides a compelling explanation for why ecologically important stimuli such as animals can be recognized in a dual-task paradigm but artificial stimuli such as a bicolor disks cannot: Through development, our visual system learns a dictionary of features that forms the basis of the position and scale tolerant representation which is found in higher level visual areas (Serre et al., 2007a). One hypothesis is that the underlying visual representation is well adapted to natural object categories but poorly adapted to artificial ones due to lack of training.

Overall, the argument above leaves open the question of the nature of the visual features that form the coarse image representation underlying rapid categorization tasks, which will be discussed in the next section. The goal of the present study is: (1) to investigate the ability of current biologically plausible models of rapid recognition to perform object recognition in natural scenes and (2) to compare the performance of these models with the state-of-the-art in computer vision as well as human observers on a rapid visual categorization task.

2. COMPUTATIONAL MODELS OF VISUAL RECOGNITION

Progress in computer vision over the past decade has been significant. Challenging visual recognition tasks such as the recognition of objects in natural scenes are no longer considered to be beyond the reach of artificial vision systems. Face detection systems are now readily available on consumer-grade digital cameras and automated face identification algorithms are being integrated in digital photo library suites. Automated pedestrian detection and computer systems for driver assistance are already available in selected vehicles, and these will become standard equipment on most models by 2014[1].

Beyond domain-specific applications, computer vision systems for the generic recognition of objects are becoming increasingly robust, as reflected by their performance on competitions such as the Pascal Challenge[2]. The number of object categories to be recognized has been increased steadily every year as the performance of the top computer vision systems has continued to improve. As it started in 2005, the challenge contained only four object categories. This year, the ImageNet Large Scale Visual Recognition Challenge[3] involved the recognition of a thousand object categories and millions of images. Overall, computer vision databases have been growing rapidly over the years with systems now being routinely tested on hundreds of object categories (Russell et al., 2007; Torralba et al., 2008, 2010; Deng et al., 2009; Everingham et al., 2010; Xiao et al., 2010).

Similarly, progress in our understanding of the computational mechanisms underlying visual recognition in cortex has been significant. Computational models have been described that have been quantitatively fitted to monkey electrophysiology data for both the processing of shape information in the ventral stream (Rust et al., 2005; David et al., 2006; Cadieu et al., 2007; Lee et al., 2007; Serre et al., 2007c; Zoccolan et al., 2007; Li et al., 2009; Cao et al., 2011; Grossberg et al., 2011; but see also Kayaert et al., 2005; Kriegeskorte et al., 2008; Op de Beeck et al., 2008; Kayaert et al., 2011) and motion information in the dorsal stream (Rust et al., 2006).

In addition, several computational models of rapid categorization have been described. **Figure 2** shows how computational models of visual recognition can be organized along three main dimensions: Feature complexity, sparse vs. dense representations, and statistical vs. non-statistical models. All the models included in this review have been shown to perform well on various visual tasks (e.g., natural scene classification, object detection, texture analysis), but the focus of the present study is limited to the ability of these models to perform well on specific visual task, namely a rapid animal categorization task (Thorpe et al., 1996). It is common to find the terms *model* and *features* used interchangeably in the literature. For clarity, in this review, we will refer to a model as a whole architecture (i.e., HMAX) and features for specific layers or components of a model (i.e., the C_1, C_2, C_{2b}, and C_3 features of the HMAX model).

2.1. FEATURE COMPLEXITY

Virtually all biological models of visual processing start by assuming an initial filtering stage. The simplest image feature that has been suggested for natural images is the WEIBULL image contrast statistics, which measures the distribution of contrast values for an image that is readily available from the response of the X and

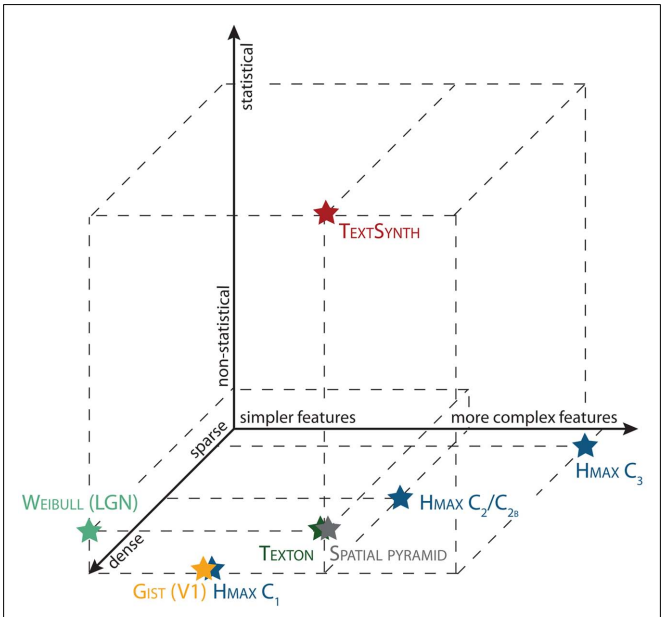

FIGURE 2 | Various computational models of visual recognition organized along three main dimensions: feature complexity, sparse vs. dense representations, and statistical vs. non-statistical models.

[1]http://www.mobileye.com.
[2]http://pascallin.ecs.soton.ac.uk/challenges/VOC
[3]www.image-net.org/challenges/LSVRC/2010

Y cells in the LGN (Ghebreab et al., 2009; Scholte et al., 2009). Weibull image contrast statistics were described as a model for natural scene identification and were shown to provide a good model of the ERP selectivity observed in EEG data.

The bottom-up Saliency algorithm by Itti and Koch (Itti et al., 1998; Itti and Koch, 2001) and the Gist algorithm by Oliva and Torralba (2001) are two examples of algorithms based on relatively low-level image features. Compared with the Weibull image contrast statistics model described above, these two models correspond to processing in the next stage of the visual hierarchy. Such models are built on the output of filter pyramids such as Gabor or steerable filters that model processing by simple and complex cells as found in the primary visual cortex (Hubel and Wiesel, 1962).

In addition to orientation, the bottom-up Saliency algorithm also includes simple feature dimensions such as contrast and color information (although only gray-value stimuli were used in the present study). While there is no *a priori* reason for the image saliency to be predictive of the presence or absence of an object category, Elazary and colleagues have shown that objects in natural scenes tend to be more salient than the background (Elazary and Itti, 2008). The performance of the Saliency model for the animal categorization task thus constitutes an interesting baseline.

Mid-level Texton features corresponding to combinations of oriented linear filter responses were shown to account for the level of performance of human observers for the classification of visual scenes (Renninger and Malik, 2004). The task tested involved the classification of natural scenes in ten categories (beach, forest, mountain, city, farm, street, bathroom, bedroom, kitchen, and living-room). Features of similar complexity were also used in the Textsynth texture synthesis algorithm by Portilla and Simoncelli (2000) and were shown to predict human performance in crowding experiments (Balas et al., 2009; see also Freeman and Simoncelli, 2011).

At the top of this hierarchy are visual features of higher complexity corresponding to multiple stages of visual processing computed by the Hmax hierarchical model of visual processing (see Serre et al., 2007a for details). Here we consider four stages of this model: (1) The C_1 stage, which corresponds to the output of V1-like oriented complex cells (and similar to the Gist features); (2) the C_2; and (3) C_{2B} stages, which have been matched to the tuning properties of cells in intermediate areas of the ventral stream of the visual cortex (area V4: Cadieu et al., 2007; Serre et al., 2007a; and PIT: Serre et al., 2007a) and correspond to features tuned to combinations of V1-like complex (C_1) units at multiple orientations, exhibiting some tolerance to changes in the position and scale of the preferred stimulus; and (4) the C_3 stage that corresponds to combinations of units from the C_2 stage. From the C_1, C_2, and C_{2B} to the C_3 layer, the visual architecture builds a hierarchical feature representation that is both increasingly complex and invariant to 2D transformations such as changes in position and scale.

To provide a baseline for the biological models, we further considered a state-of-the-art machine vision system. This popular algorithm originally introduced by Lazebnik et al. (2006) is called the Spatial Pyramid (Spatialpyr). The complexity of the features used in this algorithm is similar to the C_2/C_3 stage of the Hmax described above. The overall approach has been shown to perform well on a number of visual recognition tasks (Lazebnik et al., 2006, 2009; Bosch et al., 2007; Varma and Ray, 2007; Yang et al., 2009, 2010; Boureau et al., 2010a; Gao et al., 2010; Wang et al., 2010; Zhou et al., 2010). We believe this algorithm is representative of the current state-of-the-art in computer vision and is certainly one of the most popular.

2.2. DENSE VS. SPARSE

Another useful dichotomy between the various feature representations described above corresponds to the sparsity of the underlying visual representation. The Weibull, Gist (and C_1 stage), as well as the Texton, and the Spatialpyr are based on dense representations whereby features are matched at every location of an image.

This can be contrasted with sparse representations such as the C_{2B} and C_3 stages of the Hmax. Rather than measuring the degree of similarity between an input image and a stored representation at every position and scale, the underlying similarity in such model is based on the *best* match between a stored template and the whole image (as computed by a max operation computed across all locations and scales). Such pooling mechanisms allow the underlying representation to be tolerant to changes in position and scale.

The Textsynth algorithm probably falls somewhere in between these two extremes as it computes the statistical mean of the match across an image. For a strongly peaked distribution (as is the case for a salient object), we expect the statistical mean to closely approximate a max pooling operation and therefore behave like a sparser representation. Conversely for more textured images, one expects a broader distribution and thus the approach to behave more like a dense representation.

2.3. STATISTICAL VS. NON-STATISTICAL MODELS

Non-statistical models here refer to algorithms that are based on features computed via a simple template matching operation. Such an operation encodes the similarity between an image patch and a stored representation. In the Hmax model, an image feature at the top of the hierarchy corresponds to the best match between every patch of an input image and a stored template via a max operation.

Similarly, the Gist and the Saliency algorithms as well as some of the features of the Textsynth algorithm are based on the response of feature detectors. The Weibull image contrast statistics, the Texton algorithm, and the Spatialpyr are based on first order statistics over the computed features (i.e., histograms of the count of the index of the closest image feature over locations and scales). The Textsynth model also computes higher order statistics such as the skewness and kurtosis of the feature distributions.

3. RESULTS

3.1. MODELS PERFORMANCE IN A CATEGORIZATION TASK

Animals in natural scenes constitute a challenging class of stimuli because of the very large intra-class variations that they present. This includes large changes in appearance (terrestrial, aerial, and water animals all with a large spectrum of possible sizes) and non-rigid changes in pose, as well as clutter and changes in size and position in the visual scene. The human data presented here

appeared in a study by Serre et al. (2007b). To vary the difficulty of the task, four sets of balanced image categories were used (150 animals and 150 matching distractors in each set, i.e., 1,200 total stimuli; see Materials and Methods), each corresponding to a particular viewing distance from the camera, from an animal head to a small animal or groups of animals in cluttered natural backgrounds (see Serre et al., 2007b for details).

To minimize the role of cortical feedback by forcing visual processing to be based on a single feedforward sweep, as well as to try to minimize possible attentional shifts across the image, a backward-masking protocol (1/f dynamic noise image lasting 80 ms) was used with a long 50-ms stimulus onset asynchrony (20-ms stimulus presentation followed by a 30-ms interstimulus interval). It was found (Bacon-Macé et al., 2005) that increasing the SOA on a similar animal vs. non-animal categorization task beyond this value only has a minor effect on performance (accuracy scores for longer SOA conditions were not significantly different). At the same time, for this duration, the mask is expected to block significant feedback effects from higher level visual areas through back-projections.

Figure 3A provides an overview of the performance of the various models computed as d′ for a zero threshold value. Baseline performance by human observers (error rate calculated across all subjects; n = 22) is indicated with a black vertical line (the 95% confidence interval of the bootstrapped distribution is indicated with a gray shaded area). The estimated confidence intervals reveal that HMAX and TEXTSYNTH reached a higher level of performance than all other models (including the SPATIALPYR, a state-of-the-art computer vision system). Most importantly, the average performance of human participants fell within the confidence intervals of these two models. The performance of the remaining models decreased from the SPATIALPYR and GIST to TEXTON, WEIBULL, and SALIENCY.

The performance of individual subjects is shown on **Figure 3B** overlaid with the Receiving Operator Characteristic (ROC) curves of the computational models (corresponding to their level of performance for all possible discrimination threshold values). Each of the black dots (n = 22) corresponds to the performance of one of the human participants. The best four models (HMAX, TEXTSYNTH, SPATIALPYR, and GIST) capture relatively well the variety of behaviors exhibited by human participants: Some participants seem closer to the GIST, others to the HMAX or TEXTSYNTH). Also, it seems that the TEXTSYNTH algorithm tends to perform best at regimes with higher false alarm rates while the HMAX tends to

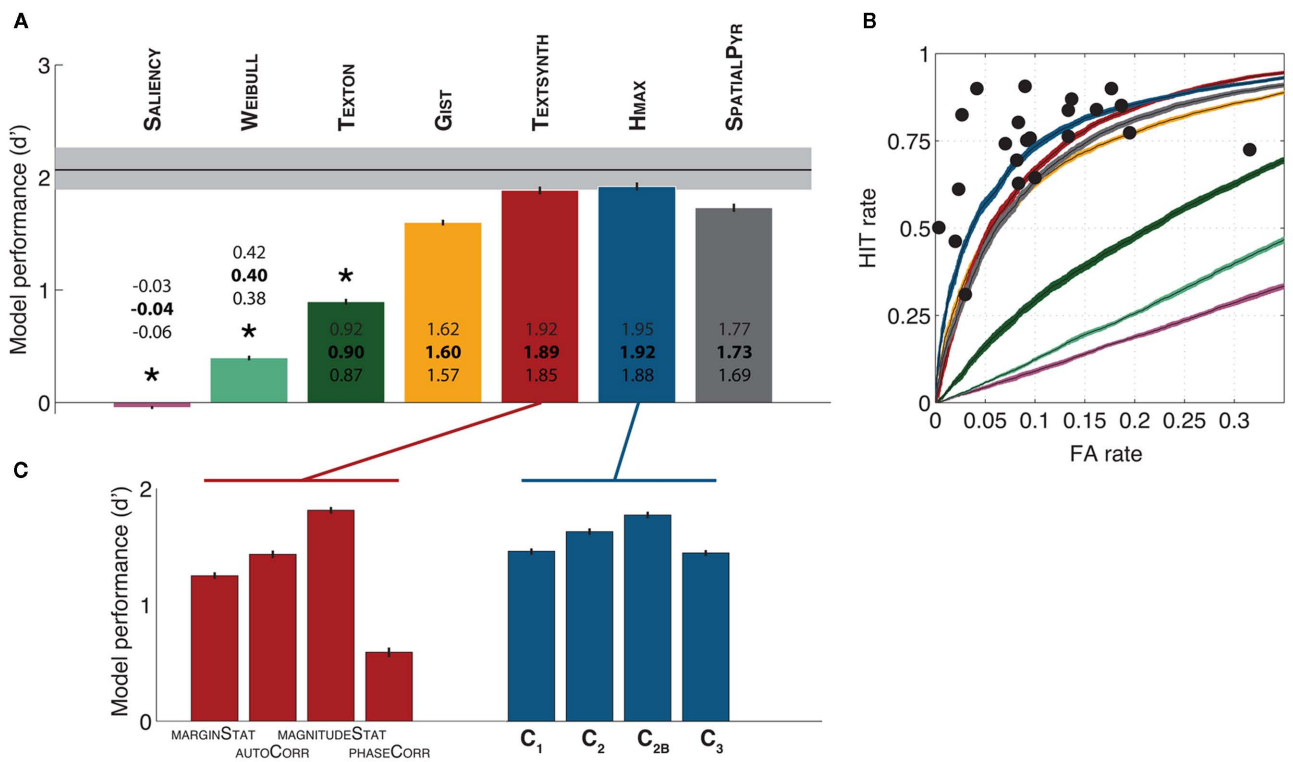

FIGURE 3 | Models performance in the animal vs. non-animal classification task. (A) Models performance measured with the d′ value (at a zero discrimination threshold value, see Materials and Methods for details). Average values (bold) as well as upper and lower limits of the corresponding confidence intervals (95% CI obtained over independent cross-validations) are reported with each bar. The performance of human observers is indicated with a black horizontal line (95% CI calculated across all subjects is indicated with a gray shaded area; see text for details). **(B)** Receiver Operating Characteristic (ROC) curves for all models tested. This shows the sensitivity (True Positive rate vs. False Alarm rate) of the different models as the discrimination threshold is being varied. Curves were averaged across multiple splits of the stimulus database used for training and testing the models. Shaded areas indicate 95% CI. Black dots correspond to the performance of individual human subjects ($n = 22$). **(C)** Performance of the various feature types in the Textsynth and Hmax models. A classifier was trained and tested independently on each feature type.

perform better at lower false alarm rates, which also seems to be the regime that best correspond to most human participants.

The HMAX and the TEXTSYNTH algorithms both rely on different types of features (see Materials and Methods). Do all features contribute equally to the reported classification results? To answer this question, we trained a classifier for each type of features separately for the two models. **Figure 3C** shows the classification performance of the resulting systems. Overall this analysis suggests that the key features are those encoding the correlations of the magnitude of responses of oriented filters for the TEXTSYNTH and the C_{2B} features for the HMAX. These two types of features indeed exhibit a similar level of complexity and tolerance to position and perform at a very similar level of performance ($d' \sim 1.8$). The C_{2B} features have been shown previously to contain a similar amount of category information (and similar tolerance to changes in position and scale) as a representative population of IT neurons (Serre et al., 2007a).

Figure 4 shows, for each model, the six images that were classified as most animal-like and most non-animal-like (as measured by the confidence of the classifiers trained on the features from

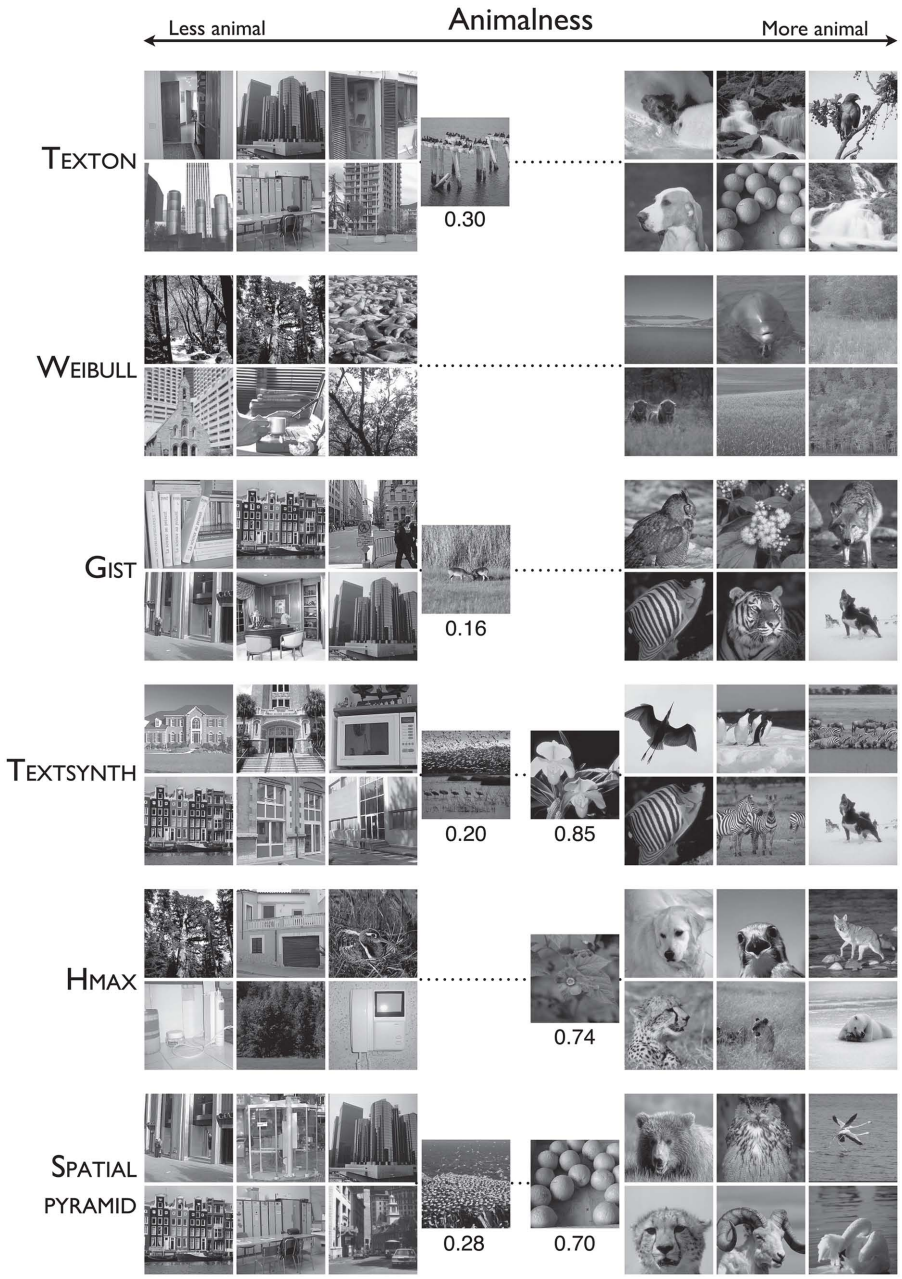

FIGURE 4 | The six most animal-like and non-animal-like images for each model. This was measured through the probability output of the classifier for each image and averaged over multiple random splits. When the six extreme images did not contain an error (e.g., an animal considered as very non-animal), the first error was added next with the corresponding score. Saliency was excluded because of its poor level of performance in the task.

each model and averaged over all random splits). From visual inspection, it appears that the most non-animal-like images for the three models that are based on first order statistics and higher (i.e., Texton, Textsynth, and the Spatialpyr) are mostly repeated textured patterns typically associated with urban scenes (e.g., buildings). Consistent with this idea, animal images that are most similar to non-animal images correspond to far groups of similar animals (e.g., flock of birds). The most animal-like images correspond to animal heads and bodies on relatively simple, near-uniform backgrounds (grass, water, snow). This is consistent with the fact that these types of features have been traditionally used for the recognition of textures (Julesz, 1981).

Interestingly the Weibull image contrast statistics seem to capture well the complexity of the stimuli with the least/most animal-like images corresponding to subjectively more complex/simpler cluttered backgrounds. However, this led to relatively poor classification performance for the present animal vs. non-animal categorization task. The behavior of the Gist and Hmax algorithms seem a bit more complex to interpret. While the Gist seems to rely heavily on the presence of vertical elements in an image, typically associated with urban and man-made scenes, for classifying an image as non-animal, no simple pattern seems to explain what constitutes representative animal-like images for either ones of the algorithms.

How do the computational models predict human performance on an individual image basis? **Figure 5** shows the correlation between the various models and human observers computed using an "animalness" score as described in Serre et al. (2007b). For human observers, this index was computed as the fraction of human observers that classified a specific image as an animal irrespective of whether its true label is animal or distractor. A score of 1.0/0.0 means that all participants classified this image as animal/non-animal. Any value in between reflects some variability across subjects.

Similarly for the computational models, a confidence score for each image was computed every time the image was selected as part of the test set (this score reflected the normalized distance to the separating decision function for this particular image on this particular split). Averaging these confidence scores across all splits resulted in one score per image that we correlated with the average scores obtained from the human observers. The bars in **Figure 5A** reflect the correlation coefficients obtained by correlating the score from human observers with the score given by each model for every image. The black horizontal bar corresponds to the inter-subject correlation between half sets of participants (procedure repeated 1,000 times with random half splits to get the 95% confidence interval indicated with the shaded area).

We found that the relative ranking of algorithms in terms of their ability to explain human performance at the single image level was indeed similar to the one based on the absolute performance as shown on **Figure 3**. To estimate how well correct classification alone for individual images impacts this score, we performed a standard permutation analysis (Good, 2000) where we shuffled all the scores randomly (diamond inset for each model) as well as a

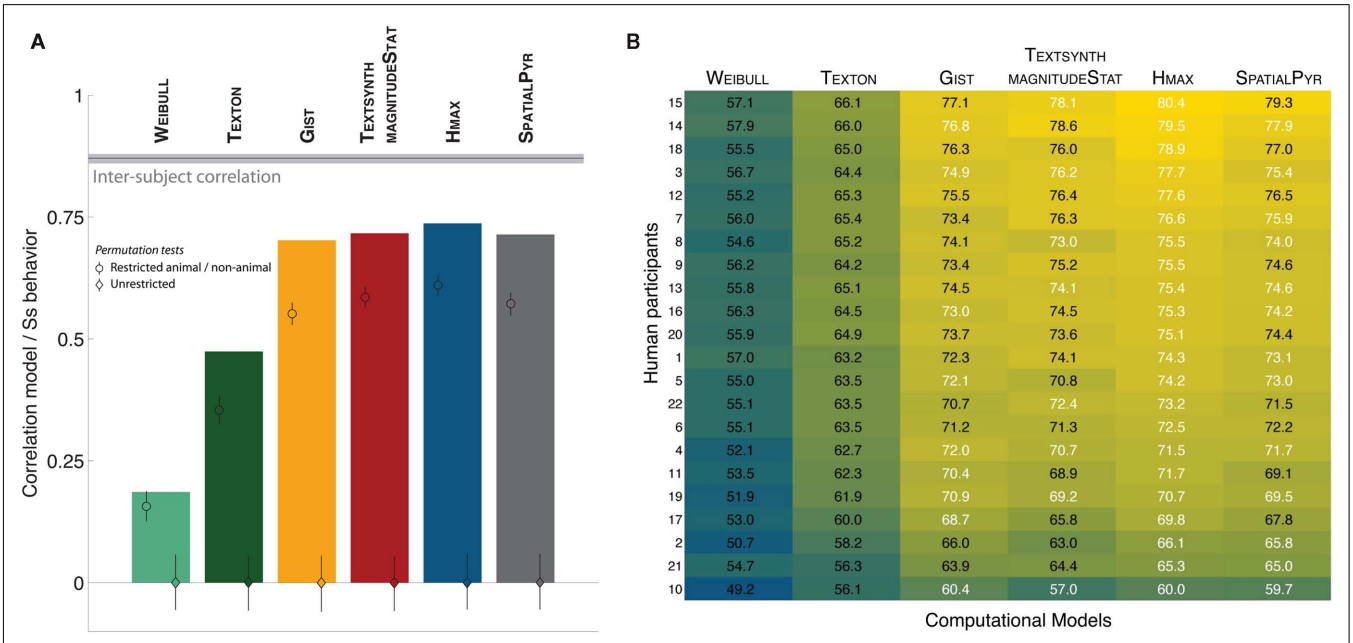

FIGURE 5 | (A) Correlation between the "animalness" scores (computed for each image) from human participants vs. computational models. For each model, the main bar corresponds to the correlation between human participants and models (a single value is computed over the whole image dataset hence no error bar shown). The diamond inserted in the bar corresponds to the correlation obtained from a standard permutation test (indicating chance level), the circle corresponds to a restricted permutation test (randomization was done for animal and non-animal stimuli separately). The most interesting difference is the one between the circle and the main bar, indicating the ability of the models to capture some of the intrinsic difficulty of individual images beyond simply classifying animal vs. non-animal images (see text for details). **(B)** Agreement matrix between each participant and the various computational models. Participants are sorted according the the highest agreement value. The correlation value for the best model and those that fall within its confidence interval are all highlighted in white.

restricted permutation procedure where we shuffled the scores for the animal and non-animal images separately (circle inset). The correlations obtained from these two restricted permutation procedures were significantly lower than those obtained for all models, except for the WEIBULL ($p = 0.058$). This suggests that with this one exception, the computational models are all able to capture some of the intrinsic difficulty of individual images. However, the correlation between even the best models and human participants remains significantly lower than the inter-subject correlation (dark horizontal bar).

We further computed the agreement between each model and individual participants. In order to obtain the matrix shown in **Figure 5B**, predicted labels from each model for every cross-validation were compared to behavioral responses from each individual participant. This allowed to get an estimate of the agreement for each participant and each computational model (over 40 cross-validations). As shown in **Figure 5B**, for every subject, HMAX was either picked as the best model for each subject or fell within the confidence interval of the actual best model when not selected as the winner (all models that fall within the confidence interval of the best model for each subject, highlighted in white). However, the GIST also seems to match at least as well or slightly better for three of the participants (subjects 4, 10, and 19). Between-participant differences could thus reflect different visual processing strategies for the task (one possibly faster but less accurate strategy based on lower-level features and one possibly slower and more robust based on higher level features).

3.2. ROBUSTNESS TO IMAGE MANIPULATIONS

Several image modification procedures are commonly used in experimental studies to assess the "high-levelness" of a visual process. Among them, the Fourier amplitude spectrum normalization and image inversion are two of the most common. It is generally believed that high-level visual processes should not be disturbed after normalization of the Fourier amplitude of the stimuli (amplitude information being low-level), but should suffer with image inversion (this modification perturbating high-level but not low-level information). It is thus interesting to test how the computational models presented here cope with these two images modifications. Considering what we know about these models and the features they extract, the results are also informative for the relevance of these image modifications in terms of how well they differentially impact high- vs. low-level visual processes.

3.2.1. *Fourier amplitude normalization*

It has been shown that the Fourier amplitude spectrum contains diagnostic information about the category of objects in natural scenes (Torralba and Oliva, 2003). Evidence regarding the use of this type of information by human participants seems restricted to specific object categories (e.g., face (VanRullen, 2006; Honey et al., 2008; Crouzet and Thorpe, 2011) but not animal category (Gaspar and Rousselet, 2009; Wichmann et al., 2010)). Computational models and the corresponding statistical classifiers are likely to take advantage of any bias in the statistics of the image sets and we thus verified whether such low-level cues could be driving the categorization performance in the previous experiment. We created a new set of images by mixing the original phase information of the original set with the averaged amplitude spectrum from all images (target and distractor images mixed as in Crouzet and Thorpe, 2011). This procedure allows to normalize the amplitude content (generally considered as lower-level visual information) while preserving the phase information (generally considered as higher-level).

The results (**Figure 6A**) first show that the performance of all the features is reduced by this image modification. However none of the features performance falls to chance level. The fact

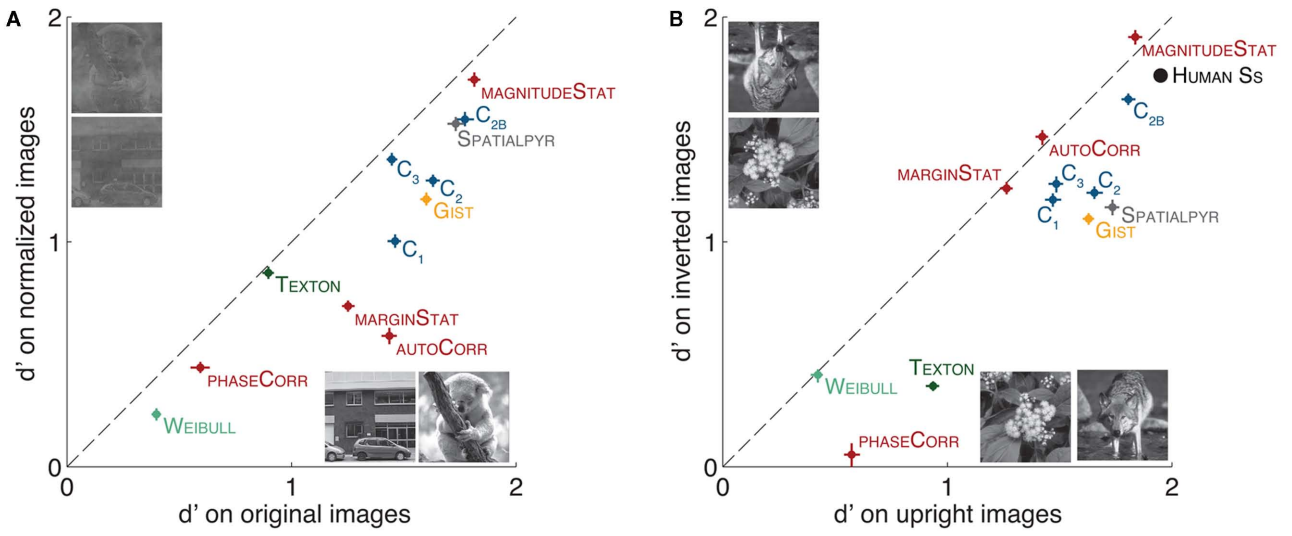

FIGURE 6 | Classification performance (d′) of the different models (features classified separately for the different components of Hmax and Textsynth) with common image modifications, compared to the performance on original images (A) Original image vs. Fourier amplitude normalized images. The normalized set was generated after averaging the amplitude spectrum content in the Fourier domain over all images (targets and distractors, a procedure identical to the one used in Crouzet and Thorpe, 2011). **(B)** Upright images vs. inverted images. Models were trained on upright images and then tested on upright (0°) or inverted (180° rotated) images. Human data from Serre et al. (2007b).

that simple low-level features like GIST, TEXTON, or the C_1 features still perform very well in the normalized condition highlights severe shortcomings associated with this procedure and its ability to completely remove low-level biases in a stimulus set. Looking more precisely at the difference between models, features like the MagnitudeStat (TEXTSYNTH) as well as the C_{2B} and C_3 (HMAX) cope very well with such image modification as observed with human participants (Gaspar and Rousselet, 2009).

3.2.2. Rotation

It is often assumed that rotating images upside-down degrades high-level information while maintaining low-level cues (which is true at least for statistics like luminance distribution and contrast). **Figure 6B** demonstrates that the effect of rotation on models performance is consistent among most of the features, irrespective of their underlying complexity. Overall we found that the observed drop in performance for most models was indeed consistent with the pattern observed for human observers in rapid categorization tasks (Rousselet et al., 2003; Guyonneau et al., 2006; Serre et al., 2007b). This suggests that perhaps this image transformation does not quite produce the effect usually intended by experimenters.

3.3. ON THE BENEFIT OF HIERARCHICAL MODELS

From the results presented above, the MAGNITUDESTAT features (as part of the TEXTSYNTH model) and the features from the higher stages of the HMAX (specifically the C_{2B} units) remain the two key contenders. As discussed above, these two types of features do indeed share some similarities as they try to capture local combinations of orientations. One key difference between these two types of features remain their invariance properties with respect to 2D transformations. While the HMAX model was designed with the goal of explaining the invariance properties of IT cells (Riesenhuber and Poggio, 1999), TEXTSYNTH was developed as a general model of texture perception without any particular focus on the problem of invariant recognition. It is thus expected that the corresponding features will exhibit significantly less tolerance to changes in position and scale.

Here to assess the invariance properties of the MAGNITUDESTAT and the C_{2B} features, we used a methodology similar to Logothetis and Pauls, 1995; see also Riesenhuber and Poggio, 1999 as well as Pinto et al., 2011 and Pinto, 2011 for a recent treatment). Here invariance is measured by first estimating a "tuning curve" (obtained by correlating a feature vector corresponding to one object at a given scale with the same feature vector obtained for the same object at different scales). An average tuning curve is then obtained by averaging tuning curves across a set of objects. Similarly a distractor response for each object is obtained by estimating the maximum correlation between the feature vector corresponding to the reference object at a given scale with the same feature vector obtained for all other objects in the set at the same scale. These responses are then averaged across all distractors and invariance to scale is then defined at the range of scales for which the correlation between the original object and its rescaled values remains significantly higher than the response to the distractors. Using 17 linearly spaced scales and 100 real-world isolated objects, **Figure 7** shows that the invariance level increases for the HMAX features throughout the hierarchy from C_1 to C_2 to C_{2B}/C_3. As seen on the Figure, the invariance properties of the C_{2B} and C_3 features remain larger than those of the TEXTSYNTH features. The fact that these two models exhibit almost identical levels of performance on the animal categorization task described above reflects a limitation of the dataset in taping in these invariant mechanisms.

As discussed in Serre and Poggio (2010), an HMAX-like representation, with built-in tolerance to position and scale of the stimulus, should, in principle, lead to a simpler classification function (such as a linear classifier as opposed to a higher order polynomial, for instance) that requires fewer training examples to achieve a specific level of performance, thus lowering the sample complexity of the recognition problem.

4. DISCUSSION

We have reviewed current computational models of rapid categorization and compared their performance to human performance on a rapid animal vs. non-animal categorization task (Serre et al., 2007b). This performance comparison revealed that HMAX and TEXTSYNTH can reach a level of performance similar to the average human observer. The GIST, which (much like the TEXTSYNTH) was not originally designed for the recognition of objects, also stands as a realistic model of visual processing and, in particular, seemed to capture well the performance of several individual participants.

Our results also suggest that features of intermediate complexity (e.g., the C_{2B} features of the HMAX model and the MAGNITUDESTAT features of the TEXTSYNTH) seem to perform better than lower-level features and on par with higher level features (the C_3 features). This is consistent with earlier proposals that features of intermediate complexity are optimal for object classification (Ullman et al., 2002). At the same time, the performance of low-level models remains relatively high on this database suggesting that despite researchers best effort to build a difficult database with changes in the position and scale of animals in these images, the dataset does exhibit some biases.

This point was already raised by Pinto et al. (2008, 2011) who showed that "natural" image databases such as the popular *CalTech-101*, because of biases for position and scale, may sometimes favor simpler models that are void of invariance properties (see also Gintautas et al., 2010; Sanbonmatsu et al., 2010). As discussed by Riesenhuber and Poggio, 1999; see also Geman, 2006), object recognition requires a difficult trade-off between invariance and selectivity. In short, more invariant features tend to be less selective and vice-versa. For instance, in Pinto et al. (2011), the C_{2B} features from the HMAX described above were shown to perform worse than V1-like features as well as other computer vision benchmarks on the CalTech-101 dataset but significantly outperform all other approaches on an artificial dataset that exhibited more variations in the position and scale of the objects thus requiring a higher level visual representation.

This idea seems consistent with the somewhat good level of performance obtained with low-level feature representations such as the C_1 features or the GIST model. Future work should address this question by selecting image subsets that are easy/difficult for lower-level vs. higher-level representations (that are tolerant to 2D transformations) and correlating the performance of human

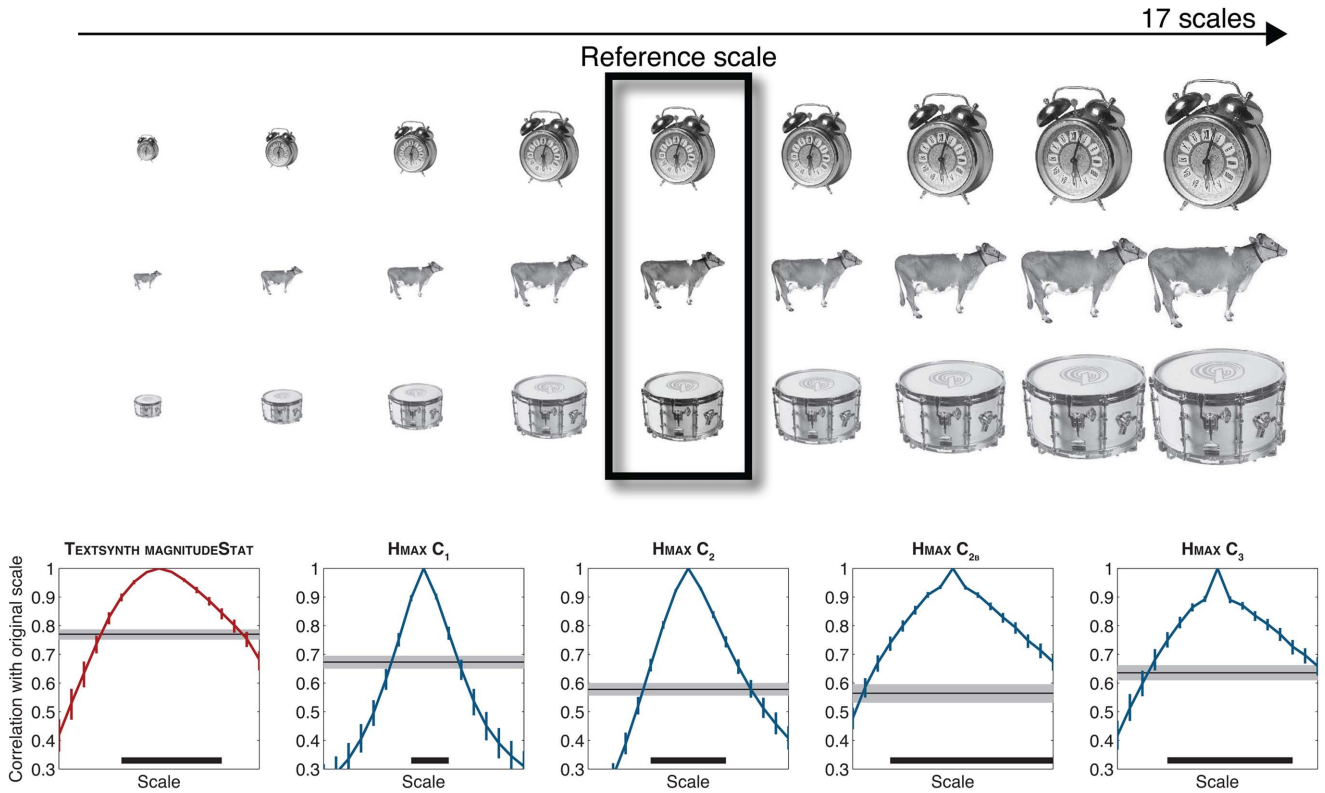

FIGURE 7 | Scale invariance in Hmax and Textsynth. "Tuning curves" were obtained by correlating feature vectors computed at one reference scale for one object from a set with feature vectors computed for the same object across multiple scales. Shown are tuning curves averaged over a set of 100 stimuli (colored curves, error bars correspond to the 95% CI). Thin black lines correspond to the average correlation for the best distractor (i.e., the distractor for which the correlation with the feature vector of the target object is highest (thin black line with 95% CI in shaded gray). The invariance properties of each model is then estimated by computing the "tuning width" such that the corresponding feature vector for a reference object remains more strongly correlated with feature vectors for the same object across scales than the best distractor at the reference scale (width displayed with the thick black line at the bottom of each graph).

observers vs. features of various levels of complexity on these subsets.

Looking at finer level correlation between computational models and human observers for individual images, we found that all tested models (with the exception of the WEIBULL) were able to capture, to some extent, some of the intrinsic difficulty of individual images beyond what can be predicted from mere performance and potentially reflect computational mechanisms similar to those used by humans. However, all models exhibited a correlation with human observers significantly lower than the inter-subject agreement, suggesting that a significant fraction of the variance in the data remains unexplained and that existing computational models do not yet fully account for the pattern of behaviors observed from human participants.

Would simple extensions of these models allow to account fully for the pattern of performance of human observers? One promising direction would involve exploring parameters (such as receptive field sizes and connectivity) by, for example, using computer-intensive parameter screening. Such an approach was shown to lead to significant improvements in classification accuracy on a face identification task using hierarchical HMAX-like/convolutional architectures (Pinto and Cox, 2011).

Another possibility could involve more efficient learning algorithms that the random sampling procedure currently implemented in the HMAX architecture. There is currently much work on the topic of feature learning both within the context of biological vision (Brumby et al., 2009) and computer vision. In computer vision, architectures related to the HMAX include deep learning networks Hinton (2010), convolutional networks (Kavukcuoglu et al., 2010; Zeiler et al., 2011), and grammar-based approaches (Fidler and Leonardis, 2007; Zhu et al., 2008). General coding strategies based on local feature pooling (Boureau et al., 2010b) and sparse coding (Mairal et al., 2008; Yang et al., 2009) constitute yet another possible avenue for improving the performance of the computational models.

Furthermore, all of the models were trained on relatively small datasets (in comparison to the number of parameters for the models) and were not explicitly optimized to match the performance of human observers (they were simply trained for the animal vs. non-animal categorization task). It is possible that higher correlation with human observers could be obtained by training these models explicitly on the pattern of responses from human observers.

Hierarchical models of the visual cortex are complex. Several layers of non-linearity coupled with the high-dimensionality of the inputs to the final classifier (due to the large number of features) makes it very hard to interpret what is driving classification accuracy in these models (Landecker et al., 2010; He et al., 2011). In particular, it has been suggested that classification of natural image categories by some of the models described above may rely more on contextual information, i.e., features computed from the background rather than the foreground (He et al., 2011). These types of behavior most likely results from the relative small number of images used to train and the comparatively high-dimensionality of the feature vectors used for classification.

He et al. (2011) described a hierarchical (probabilistic) model whereby natural object categories are represented by a coarse hierarchical probability distribution over object geometry and spatial configuration of object parts. Because of this and the need for (manual) segmentation of animal images for training the model suggests that this class of models might however require attentional mechanisms and cortical feedback. While it might be incompatible with the severe time constraints imposed by rapid categorization tasks, such models do however suggest possible avenues for improving the performance of the computational models beyond the first initial feedforward sweep. Similarly Chikkerur et al. (2010) have shown that an extension of the HMAX model with feature-based and spatial attention was able to further improve recognition performance of the model on the task.

However, it is important to realize the intrinsic limitations of the specific computational framework we have described and why it is at best a first step toward understanding the visual cortex. Some important limitations for these types of object representations based on a loose collection of image features is that they remain sensitive to the presence of background clutter (Serre et al., 2007b; Chikkerur et al., 2010), do not explicitly encode spatial relations between parts that are known to play a key role in object recognition (Biederman, 1987) and do not explicitly distinguish figure from ground (Lamme and Roelfsema, 2000). Most importantly, given enough time, humans use eye movement to scan images, and performance in many object recognition tasks improves significantly over that obtained during quick presentations.

While these models remain simplistic models of visual processing, they do suggest an alternative to the classical visual pipeline sketched on **Figure 1** (left), which places an emphasis on bottom-up computations for grouping, Figure-ground segmentation, and spatial relations. Instead this alternative view suggests that the bottom-up activation of a loose collection of hardwired feature detectors via a hierarchy of increasing complex processing stages may provide a coarse initial visual representation for more complex routines and several feedforward/feedback iterations to solve specific tasks including edge detection, grouping, figure segregation, and the computation of spatial relations between parts, among others, and more generally the parsing and interpretation of complex visual scenes (see for instance, Hochstein and Ahissar, 2002; Bar, 2004; Zheng et al., 2007; Epshtein et al., 2008; Serre and Poggio, 2010 for a recent review).

5. MATERIALS AND METHODS

5.1. COMPUTATIONAL MODELS

Below we describe in greater detail the models used in this study. Most of the softwares were publicly available from the authors web sites and/or provided by the authors. We tried to equalize as much as possible the various model parameters when possible (e.g., number of frequency bands and orientations, etc.,). When the benefit on performance was not significant, default parameters were kept.

Saliency

Models of bottom-up saliency compute the local conspicuity of an image region with respect to its surround as measured, for instance, by local contrast, color, or orientation. Here we used the matlab Saliency Toolbox 2.1 (Walther and Koch, 2006). Low-level features (pixel intensity, orientations) were extracted at multiple scales and local conspicuity maps were computed using local center-surround mechanisms. Note that color was not used here because the stimuli used were grayscale. The resulting conspicuity maps were then combined to form a saliency map to predict the location of the highest saliency value for the whole image (Itti et al., 1998; Itti and Koch, 2001). This intensity value (single feature) was then used to try to predict the presence or absence of an animal in images.

Weibull

The distribution of local contrast in an image can be well fitted with the so-called WEIBULL distribution (Ghebreab et al., 2009; Scholte et al., 2009). Such distribution can be estimated from the output of zero-crossing detectors similar to the center-surround cells found in the LGN (Scholte et al., 2009). These authors further hypothesized that this information could be available very rapidly for the visual system and as such be used for rapid categorization. Indeed, they showed that the β and γ parameters of the WEIBULL distribution correlate well with EEG activity and could even allow identification of the precise image presented to a human subject among a small set of natural images (Ghebreab et al., 2009). More precisely, the β and γ parameters of the WEIBULL distribution define a space where images are ordered according to their level of clutter/complexity and texture similarity (Scholte et al., 2009). Here we used the code provided by the authors which uses the simple β and γ parameters from the fitted WEIBULL distribution. We found the performance of the two models presented in Scholte et al. (2009) to be very similar and only report here the performance of the simpler abstract one.

Gist

The GIST features correspond to the model by Oliva & Torralba, who have shown that the amplitude spectrum of images in the Fourier domain could be predictive of scene category, leading later to the concept of spatial envelope or global image signature (Oliva and Torralba, 2001, 2006; Torralba and Oliva, 2003). To create this representation, global features (Torralba and Oliva, 2003) were computed by convolving each image in the database with a Gabor filter pyramid (8 levels and 8 orientations) and further down-sampling the resulting filtered image to produce a $4 \times 4 \times 64$ (=1,024) dimensional vector, which is then used for classification.

Texton

The Texton features were described in Renninger and Malik (2004). They were computed with a filter pyramid (M = 96 Gabor filters at 8 orientations, 6 scales, and 2 phases). A large number of random patches were extracted from hundreds of M-dimensional edge-response images (from the pre-training set) and subsequently clustered using k-means to find 100 cluster centroids. Each of these centroids then became a Texton feature. For every image, an image of Texton counts was computed by finding the index of the nearest Texton to the vector of filter responses at each pixel location and accumulating the counts over the whole image leading to 100-dimensional histogram vector used for classification as done in Renninger and Malik (2004).

TextSynth

Textsynth was originally presented as a model of parametric texture analysis/synthesis[4] (Portilla and Simoncelli, 2000). Its success at producing new texture images from random noise that seem to be perceptually similar to a seed image makes it an interesting addition to our study. Recent work by Balas et al. (2009) and Freeman and Simoncelli (2011) suggests that the model accounts well for the representation of the early visual system, and in particular can stand as a good model for crowding in the visual periphery if filter size is made larger in the periphery.

This model first measures the responses of oriented linear filters, which are computed using a complex-valued steerable pyramid decomposition. This approximates the response of V1 complex cells tuned to different orientations and scales (4 orientations and 4 scales, higher values for these parameters were tested but did not improve performance) which tile all positions in the image. Then, the model computes joint statistics of these features to capture intermediate-level image structure. The statistics used in this model fall into four main categories: (1) Marginal statistics: the marginal distribution of luminance in the image (i.e., mean, variance, skew, and kurtosis as well as skewness and kurtosis of the low-pass image); (2) Luminance autocorrelation (as a proxy for the detection of periodic structures in the stimulus); (3) Magnitude statistics: the correlations of the magnitude of the responses of oriented wavelets across differences in orientation, neighboring positions, and scales (to capture simple structures in the image such as lines, edges, corners, and junctions); and (4) Phase statistics: phase correlation across scales (in order to capture the alignment of phase structure in local features).

Hmax

The Hmax model of object recognition combines a hierarchical build-up of invariance and complexity (inspired by Fukushima, 1980) with the idea of view-based recognition of 3D objects (Riesenhuber and Poggio, 1999, 2000). Here we used the extended model described by Serre et al. (2007b,c). Over the years, several related hierarchical models have been developed (Mel, 1997; Wallis and Rolls, 1997; LeCun et al., 1998; Riesenhuber and Poggio, 1999; Ullman et al., 2002; Amit and Mascaro, 2003; Wersing and Koerner, 2003; Masquelier and Thorpe, 2007; Mutch and Lowe, 2008; Jarrett et al., 2009; Pinto et al., 2011; Saxe et al., 2011). We here focus on the Hmax because the underlying parameters of the architecture were explicitly derived from available neuroscience data. This system-level computer model seems consistent with monkey electrophysiological data in different cortical areas of the ventral visual pathway (Serre et al., 2007a) as well as human behavioral data during rapid categorization tasks with natural images (Serre et al., 2007b). These findings suggested that bottom-up processes may provide a satisfactory description of the very first pass of information in the visual cortex.

Here, we used the GPU implementation developed by Mutch et al. (2010), with the default parameters of the Hmax demo with the exception of the number of orientations and scales that were set to 8 (to better match the parameters of the other models). The dictionary was learned/extracted on a pre-training set of 128 natural images (different from the ones used for the training and test of the recognition stage). Here we used 2,048 C_1 and C_2 features selected at random. All 2,048 C_{2B} and 1,024 C_3 features were used.

SpatialPyr

This state-of-the-art computer vision system (Lazebnik et al., 2006; Yang et al., 2009, 2010; Boureau et al., 2010a; Gao et al., 2010; Wang et al., 2010; Zhou et al., 2010) provides a useful baseline for the biologically inspired models described above. The approach is based on increasingly fine sub-divisions of an image into partitions and the computation of local features histograms within these subregions. The resulting spatial pyramid has been shown to provide a simple and computationally efficient representation as demonstrated by the high-level of performance of the system on several image classification tasks (Lazebnik et al., 2006; Yang et al., 2009, 2010; Boureau et al., 2010a; Gao et al., 2010; Wang et al., 2010; Zhou et al., 2010).

5.2. IMAGE DATABASE

Here we consider the animal and non-animal dataset from the study by Serre et al. (2007a). The dataset contains 1,200 images (600 animals and 600 non-animals). As a pre-processing step, images were converted to grayscale and resized to be 256 × 256. Two of the computational models tested (Texton and Hmax) required the learning of a dictionary of features. As done in Serre et al. (2007a), we considered an additional set of 128 natural images containing various categories (from animals and vehicles to landscapes and human faces) specifically for the extraction of features and the learning of codebooks in these two models.

5.3. HUMAN DATA

Here we compare the models presented above to the human behavioral data collected by Serre et al. (2007b). In this rapid categorization task, the images were flashed for 20 ms on the screen, followed by a blank screen for 30 ms, and then a mask appeared for 80 ms (the Stimulus Onset Asynchrony was thus 50 ms). The participants had to respond as quickly as possible, indicating whether they saw an animal or a distractor image by pressing one of two keys (see Serre et al., 2007b) for a more detailed description. A total of twenty-four human observers participated in the original study. Two of the participants were excluded because of their overall poor level of performance.

[4]http://www.cns.nyu.edu/lcv/texture/

5.4 CLASSIFICATION

To assess the diagnosticity of the various visual features described above for the categorization of animal and non-animal images, a linear Support Vector Machine (SVM) classifier was used (Fan et al., 2008, LIBLINEAR 1.7). The procedure runs as follow: First, the 1,200 image from the animal/non-animal database were equally split in a training set and a test set that contains an equal proportion of target (300) and distractor images (300). Second, an optimal cost parameter C was determined through line search optimization using 8-fold cross-validation on the training set of images. An SVM classifier was then trained and tested on the various types of features (the exact number of features used depended on the type of models considered, see above). For each model, the reported results correspond to the average performance (and corresponding 95% confidence intervals) using a cross-validation procedure ($n = 40$) whereby different training and test sets were selected each time at random.

5.5 INVARIANCE TEST

Here we considered the database of images used in Konkle and Oliva (2011) containing 100 isolated real-world objects. Features were extracted from all images at seventeen different scales and the middle scale was selected as reference. For each object and computational model, we computed a "tuning curve" based on the correlation between the feature vector obtained for the reference scale and the feature vector obtained for all remaining scales for the same object. We also computed a "distractor curve" based on the maximum correlation between the feature vector obtained for the reference object at the reference scale with all remaining (distractor) objects at the same reference scale (see Logothetis and Pauls, 1995 for details). A t-test (Two-sample, False Detection Rate correction for multiple comparison, α corrected from 0.05 to 0.024) was performed at every scale in order to compare the values obtained for the "tuning curve" and for the "distractor curve." This allows to obtain the "tuning width" of the representation (thick black bar on **Figure 7**).

ACKNOWLEDGMENTS

The authors would like to thank the original authors of the computational models for providing the code or having made their code available on the internet. This work was supported by DARPA (DARPA-BAA-09-31). Part of this research was conducted using computational resources and services at the Center for Computation and Visualization, Brown University.

REFERENCES

Amit, Y., and Mascaro, M. (2003). An integrated network for invariant visual detection and recognition. *Vision Res.* 43, 2073–2088.

Bacon-Macé, N., Macé, M. J.-M., Fabre-Thorpe, M., and Thorpe, S. J. (2005). The time course of visual processing: backward masking and natural scene categorisation. *Vision Res.* 45, 1459–1469.

Balas, B., Nakano, L., and Rosenholtz, R. (2009). A summary-statistic representation in peripheral vision explains visual crowding. *J. Vis.* 9, 1–18.

Bar, M. (2004). Visual objects in context. *Nat. Rev. Neurosci.* 5, 617–629.

Bergen, J. R., and Julesz, B. (1983). Parallel versus serial processing in rapid pattern discrimination. *Nature* 303, 696–698.

Biederman, I. (1987). Recognition-by-components: a theory of human image understanding. *Psychol. Rev.* 94, 115–147.

Bosch, A., Zisserman, A., and Munoz, X. (2007). "Representing shape with a spatial pyramid kernel," in *Proceedings of the 6th ACM International Conference on Image and Video Retrieval*, Amsterdam, 401–408.

Boureau, Y.-L., Bach, F., LeCun, Y., and Ponce, J. (2010a). "Learning mid-level features for recognition," in *Computer Vision and Pattern Recognition (CVPR)* (IEEE), San Francisco, CA.

Boureau, Y.-L., Ponce, J., and LeCun, Y. (2010b). "A theoretical analysis of feature pooling in visual recognition," in *27th International Conference on Machine Learning* (HAIFA: Citeseer).

Brumby, S. P., Kenyon, G., Landecker, W., Rasmussen, C., Swaminarayan, S., and Bettencourt, L. M. A. (2009). "Large-scale functional models of visual cortex for remote sensing," in *IEEE Applied Imagery Pattern Recognition Workshop* (IEEE), Washington, DC, 1–6.

Cadieu, C., Kouh, M., Pasupathy, A., Connor, C., Riesenhuber, M., and Poggio, T. A. (2007). A model of V4 shape selectivity and invariance. *J. Neurophysiol.* 98, 1733–1750.

Cao, Y., Grossberg, S., and Markowitz, J. (2011). How does the brain rapidly learn and reorganize view- and positionally-invariant object representations in inferior temporal cortex? *Neural Netw.* 24, 1050–1061.

Chikkerur, S. S., Serre, T., Tan, C., and Poggio, T. (2010). What and where: a bayesian inference theory of attention. *Vision Res.* 50, 2233–2247.

Crouzet, S. M., Kirchner, H., and Thorpe, S. J. (2010). Fast saccades towards face: face detection in just 100 ms. *J. Vis.* 10, 1–17.

Crouzet, S. M., and Thorpe, S. J. (2011). Low level cues and ultra-fast face detection. *Front. Psychology* 2:342. doi: 10.3389/fpsyg.2011.00342

David, S. V., Hayden, B. Y., and Gallant, J. L. (2006). Spectral receptive field properties explain shape selectivity in area V4. *J. Neurophysiol.* 96, 3492–3505.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). "ImageNet: a large-scale hierarchical image database," in *Proceedings of the Conference on Computer Vision and Pattern Recognition*, Miami, FL, 248–255.

Elazary, L., and Itti, L. (2008). Interesting objects are visually salient. *J. Vis.* 8, 1–15.

Epshtein, B., Lifshitz, I., and Ullman, S. (2008). Image interpretation by a single bottom-up top-down cycle. *Proc. Natl. Acad. Sci. U.S.A.* 105, 14298–14303.

Evans, K. K., and Treisman, A. (2005). Perception of objects in natural scenes: is it really attention free? *J. Exp. Psychol. Hum. Percept. Perform.* 31, 1476–1492.

Everingham, M., Van Gool, L., Williams, C., Winn, J., and Zisserman, A. (2010). The PASCAL visual object classes (VOC) challenge. *Int. J. Comput. Vis.* 88, 303–338.

Fan, R.-E., Wang, X.-R., and Lin, C.-J. (2008). LIBLINEAR: a library for large linear classification. *J. Mach. Learn. Res.* 9, 1871–1874.

Fidler, S., and Leonardis, A. (2007). "Towards scalable representations of object categories: learning a hierarchy of parts," in *IEEE Conference on Computer Vision and Pattern Recognition* (IEEE), Minneapolis, MN, 1–8.

Freeman, J., and Simoncelli, E. (2011). Metamers of the ventral stream. *Nat. Neurosci.* 14, 1195–1201.

Fukushima, K. (1980). Neocognitron: a self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biol. Cybern.* 36, 193–202.

Gao, S., Tsang, I., Chia, L., and Zhao, P. (2010). "Local features are not lonely? Laplacian sparse coding for image classification," in *Computer Vision and Pattern Recognition (CVPR)* (IEEE), San Francisco, CA.

Gaspar, C. M., and Rousselet, G. A. (2009). How do amplitude spectra influence rapid animal detection? *Vis. Res.* 49, 3001–3012.

Geman, S. (2006). Invariance and selectivity in the ventral visual pathway. *J. Physiol. Paris* 100, 212–224.

Ghebreab, S., Scholte, S., Lamme, V. A. F., and Smeulders, A. (2009). "A biologically plausible model for rapid natural scene identification," in *Advances in Neural Information Processing Systems*, eds Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, 22, 629–637.

Gintautas, V., Kunsberg, B., Ham, M., Barr, S., Zucker, S., Brumby, S., Bettencourt, L. M. A., and Kenyon, G. T. (2010). An improved model for contour completion in V1 using learned feature correlation statistics. *J. Vis.* 10, 1162.

Good, P. (2000). *Permutation Tests: A Practical Guide to Resampling Methods for Testing Hypotheses.* New York, NY: Springer.

Grossberg, S., Markowitz, J., and Cao, Y. (2011). On the road to invariant recognition: explaining tradeoff and morph properties of cells in inferotemporal cortex using multiple-scale task-sensitive attentive learning. *Neural Netw.* 24, 1036–1049.

Guyonneau, R., Kirchner, H., and Thorpe, S. J. (2006). Animals roll around the clock: the rotation invariance of ultrarapid visual processing. *J. Vis.* 6, 1008–1017.

He, X., Yang, Z., and Tsien, J. Z. (2011). A hierarchical probabilistic model for rapid object categorization in natural scenes. *PLoS ONE* 6, e20002. doi:10.1371/journal.pone.0020002

Hinton, G. E. (2010). Learning to represent visual input. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 365, 177–184.

Hochstein, S., and Ahissar, M. (2002). View from the top: hierarchies and reverse hierarchies in the visual system. *Neuron* 36, 791–804.

Honey, C., Kirchner, H., and VanRullen, R. (2008). Faces in the cloud: Fourier power spectrum biases ultrarapid face detection. *J. Vis.* 8, 9.

Hubel, D., and Wiesel, T. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *J. Physiol.* 160, 106–154.

Hubel, D., and Wiesel, T. N. (1959). Receptive fields of single neurones in the cat's striate cortex. *J. Physiol. (Lond.)* 148, 574–591.

Hubel, D., and Wiesel, T. N. (1968). Receptive fields and functional architecture of monkey striate cortex. *J. Physiol. (Lond.)* 195, 215–243.

Itti, L., and Koch, C. (2001). Computational modelling of visual attention. *Nat. Rev. Neurosci.* 2, 194–203.

Itti, L., Koch, C., and Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Trans. Pattern Anal. Mach. Intell.* 20, 1254–1259.

Jarrett, K., Kavukcuoglu, K., Ranzato, M. A., and LeCun, Y. (2009). "What is the best multi-stage architecture for object recognition?" *2009 IEEE 12th International Conference on Computer Vision*, Kyoto, 2146–2153.

Julesz, B. (1981). Textons, the elements of texture perception, and their interactions. *Nature* 290, 91–97.

Kavukcuoglu, K., Sermanet, P., Boureau, Y.-L., Gregor, K., Mathieu, M., and LeCun,Y. (2010). "Learning convolutional feature hierarchies for visual recognition," in *Advances in Neural Information Processing Systems*, eds J. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, and A. Culotta, 23, 1090–1098.

Kayaert, G., Biederman, I., and Vogels, R. (2005). Representation of regular and irregular shapes in macaque inferotemporal cortex. *Cereb. Cortex* 15, 1308–1321.

Kayaert, G., Wagemans, J., and Vogels, R. (2011). Encoding of complexity, shape, and curvature by macaque infero-temporal neurons. *Front. Syst. Neurosci.* 5:51. doi:10.3389/fnsys.2011.00051

Kirchner, H., and Thorpe, S. J. (2006). Ultra-rapid object detection with saccadic eye movements: Visual processing speed revisited. *Vis. Res.* 46, 1762–1776.

Konkle, T., and Oliva, A. (2011). Canonical visual size for real-world objects. *J. Exp. Psychol. Hum. Percept. Perform.* 37, 23–37.

Kriegeskorte, N., Mur, M., and Bandettini, P. (2008). Representational similarity analysis – connecting the branches of systems neuroscience. *Front. Syst. Neurosci.* 2:4. doi:10.3389/neuro.06.004.2008

Lamme, V. A. F., and Roelfsema, P. (2000). The distinct modes of vision offered by feedforward and recurrent processing. *Trends Neurosci.* 23, 571–579.

Landecker, W., Brumby, S., Thomure, M., Kenyon, G., Bettencourt, L., and Mitchell, M. (2010). "Visualizing classification decisions of hierarchical models of cortex," in *Computational and Systems Neuroscience (COSYNE)*, Salt Lake City, UT.

Lazebnik, S., Schmid, C., and Ponce, J. (2006). Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. *IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.* 2, 2169–2178.

Lazebnik, S., Schmid, C., and Ponce, J. (2009). "Spatial pyramid matching," in *Object Categorization: Computer and Human Vision Perspectives*, Citeseer, 401–415.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proc. IEEE* 86, 2278–2324.

Lee, H., Ekanadham, C., and Ng, A. (2007). Sparse deep belief net model for visual area V2. *Adv. Neural Inf. Process Syst.* 19, 1–8.

Li, F., VanRullen, R., Koch, C., and Perona, P. (2002). Rapid natural scene categorization in the near absence of attention. *Proc. Natl. Acad. Sci. U.S.A.* 99, 9596.

Li, N., Cox, D. D., Zoccolan, D., and DiCarlo, J. J. (2009). What response properties do individual neurons need to underlie position and clutter "invariant" object recognition? *J. Neurophysiol.* 102, 360–376.

Logothetis, N. K., and Pauls, J. (1995). Psychophysical and physiological evidence for viewer-centered object representations in the primate. *Cereb. Cortex* 5, 270–288.

Mairal, J., Bach, F., Ponce, J., Sapiro, G., and Zisserman, A. (2008). "Discriminative learned dictionaries for local image analysis," in 2008 *IEEE Conference on Computer Vision and Pattern Recognition*, Anchorage, AK, 1–8.

Marr, D. (1982). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information.* San Francisco: W. H. Freeman & Co Ltd.

Masquelier, T., and Thorpe, S. J. (2007). Unsupervised learning of visual features through spike timing dependent plasticity. *PLoS Comput. Biol.* 3, e31. doi:10.1371/journal.pcbi.0030031

Mel, B. W. (1997). SEEMORE: combining color, shape, and texture histogramming in a neurally inspired approach to visual object recognition. *Neural Comput.* 9, 777–804.

Mutch, J., Knoblich, U., and Poggio, T. (2010). *CNS: A GPU-Based Framework for Simulating Cortically-Organized Networks.* Technical report, MIT-CSAIL-TR-2010-013/CBCL-286. Cambridge, MA: Massachusetts Institute of Technology.

Mutch, J., and Lowe, D. G. (2008). Object class recognition and localization using sparse features with limited receptive fields. *Int. J. Comput. Vis.* 80, 45–57.

Oliva, A., and Torralba, A. (2001). Modeling the shape of the scene: a holistic representation of the spatial envelope. *Int. J. Comput. Vis.* 42, 145–175.

Oliva, A., and Torralba, A. (2006). Building the gist of a scene: the role of global image features in recognition. *Prog. Brain Res.* 155, 23.

Op de Beeck, H. P., Torfs, K., and Wagemans, J. (2008). Perceived shape similarity among unfamiliar objects and the organization of the human object vision pathway. *J. Neurosci.* 28, 10111–10123.

Pinto, N. (2011). *Forward Engineering Object Recognition: A Scalable Approach.* Ph.D. thesis, Massachusetts Institute of Technology, Boston, CA.

Pinto, N., Barhomi, Y., Cox, D. D., and Dicarlo, J. J. (2011). "Comparing state-of-the-art visual features on invariant object recognition tasks," in *IEEE Workshop on Applications of Computer Vision (WACV)*, Kona, HI.

Pinto, N., and Cox, D. D. (2011). "Beyond simple features: a large-scale feature search approach to unconstrained face recognition," in *IEEE International Conference on Automatic Face and Gesture Recognition*, Santa Barbara, CA.

Pinto, N., Cox, D. D., and DiCarlo, J. J. (2008). Why is real-world visual object recognition hard. *PLoS Comput. Biol.* 4, e27. doi:10.1371/journal.pcbi.0040027

Portilla, J., and Simoncelli, E. P. (2000). A parametric texture model based on joint statistics of complex wavelet coefficients. *Int. J. Comput. Vis.* 40, 49–71.

Potter, M. C., and Levy, E. I. (1969). Recognition memory for a rapid sequence of pictures. *J. Exp. Psychol. Hum. Percept. Perform.* 81, 10–15.

Renninger, L., and Malik, J. (2004). When is scene identification just texture recognition? *Vision Res.* 44, 2301–2311.

Riesenhuber, M., and Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nat. Neurosci.* 2, 1019–1025.

Riesenhuber, M., and Poggio, T. (2000). Models of object recognition. *Nat. Neurosci.* 3, 1199–1204.

Rousselet, G. A., Fabre-Thorpe, M., and Thorpe, S. J. (2002). Parallel processing in high-level categorization of natural images. *Nat. Neurosci.* 5, 629–630.

Rousselet, G. A., Macé, M. J.-M., and Fabre-Thorpe, M. (2003). Is it an animal? Is it a human face? Fast processing in upright and inverted natural scenes. *J. Vis.* 3, 440–455.

Russell, B. C., Torralba, A., Murphy, K. P., and Freeman, W. T. (2007). LabelMe: a database and web-based tool for image annotation. *Int. J. Comput. Vis.* 77, 157–173.

Rust, N. C., Mante, V., Simoncelli, E. P., and Movshon, J. A. (2006). How MT cells analyze the motion of visual patterns. *Nat. Neurosci.* 9, 1421–1431.

Rust, N. C., Schwartz, O., Movshon, J. A., and Simoncelli, E. P. (2005). Spatiotemporal elements of macaque v1 receptive fields. *Neuron* 46, 945–956.

Sanbonmatsu, K., Bennett, R., Barr, S., Renaudo, C., Ham, M., Gintautas, V., Brumby, S., George, J., Kenyon, G., and Bettencourt, L. (2010). Comparing speed-of-sight studies using rendered vs. natural images. *J. Vis.* 10, 986.

Saxe, A., Koh, P., Chen, Z., Bhand, M., Suresh, B., and Ng, A. (2011). "On random weights and unsupervised feature learning," in *Twenty-Eighth International Conference on Machine Learning*, Vancouver, BC, 1–9.

Scholte, H., Ghebreab, S., Waldorp, L., Smeulders, A., and Lamme, V. A. F. (2009). Brain responses strongly correlate with Weibull image statistics when processing natural images. *J. Vis.* 9, 1–15.

Serre, T., Kreiman, G., Kouh, M., Cadieu, C., Knoblich, U., and Poggio, T. (2007a). A quantitative theory of immediate visual recognition. *Prog. Brain Res.* 165, 33.

Serre, T., Oliva, A., and Poggio, T. (2007b). A feedforward architecture accounts for rapid categorization. *Proc. Natl. Acad. Sci. U.S.A.* 104, 6424–6429.

Serre, T., Wolf, L., Bileschi, S., Riesenhuber, M., and Poggio, T. (2007c). Robust object recognition with cortex-like mechanisms. *IEEE Trans. Pattern Anal. Mach. Intell.* 29, 411–426.

Serre, T., and Poggio, T. (2010). A neuromorphic approach to computer vision. *Commun. ACM* 53, 54.

Thorpe, S. J., Fize, D., and Marlot, C. (1996). Speed of processing in the human visual system. *Nature* 381, 520–522.

Torralba, A., Fergus, R., and Freeman, W. T. (2008). 80 million tiny images: a large data set for nonparametric object and scene recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* 30, 1958–1970.

Torralba, A., and Oliva, A. (2003). Statistics of natural image categories. *Network* 14, 391–412.

Torralba, A., Russell, B. C., and Yuen, J. (2010). LabelMe: online image annotation and applications. *Proc. IEEE* 98, 1467–1484.

Treisman, A., and Gelade, G. (1980). A feature-integration theory of attention. *Cogn. Psychol.* 136, 97–136.

Ullman, S. (1996). *High-Level Vision: Object Recognition and Visual Cognition.* Cambridge, MA: The MIT Press.

Ullman, S., Vidal-Naquet, M., and Sali, E. (2002). Visual features of intermediate complexity and their use in classification. *Nat. Neurosci.* 5, 682–687.

VanRullen, R. (2006). On second glance: Still no high-level pop-out effect for faces. *Vision Res.* 46, 3017–3027.

VanRullen, R. (2007). The power of the feed-forward sweep. *Adv. Cogn. Psychol.* 3, 167–176.

VanRullen, R., and Koch, C. (2003). Visual selective behavior can be triggered by a feed-forward process. *J. Cogn. Neurosci.* 15, 209–217.

VanRullen, R., and Thorpe, S. J. (2001). The time course of visual processing: from early perception to decision-making. *J. Cogn. Neurosci.* 13, 454–461.

Varma, M., and Ray, D. (2007). "Learning the discriminative power-invariance trade-off," in *International Conference on Computer Vision (ICCV)* (IEEE), Rio de Janeiro, 1–8.

Wallis, G., and Rolls, E. T. (1997). A model of invariant object recognition in the visual system. *Prog. Neurobiol.* 51, 167–194.

Walther, D., and Koch, C. (2006). Modeling attention to salient proto-objects. *Neural Netw.* 19, 1395–1407.

Wang, J., Yang, J., Yu, K., Lv, F., Huang, T., and Gong, Y. (2010). "Locality-constrained linear coding for image classification," in *Computer Vision and Pattern Recognition (CVPR)*, IEEE Conference, San Francisco, CA, 3360–3367.

Wersing, H., and Koerner, E. (2003). Learning optimized features for hierarchical models of invariant recognition. *Neural Comput.* 15, 1559–1588.

Wichmann, F. A., Drewes, J., Rosas, P., and Gegenfurtner, K. R. (2010). Animal detection in natural scenes: critical features revisited. *J. Vis.* 10, 1–27.

Wolfe, J. M. (2006). "Guided search 4.0," in *Integrated Models of Cognitive Systems,* ed. W. Gray (New York: Oxford), 99–120.

Wolfe, J. M., and Horowitz, T. S. (2004). What attributes guide the deployment of visual attention and how do they do it? *Nat. Rev. Neurosci.* 5, 495–501.

Xiao, J., Hays, J., Ehinger, K., Oliva, A., and Torralba, A. (2010). "SUN database: large-scale scene recognition from abbey to zoo," in *Computer Vision and Pattern Recognition (CVPR)*, IEEE Conference, San Francisco, CA, 3485–3492.

Yang, J., Yu, K., Gong, Y., and Huang, T. (2009). "Linear spatial pyramid matching using sparse coding for image classification," in *Computer Vision and Pattern Recognition (CVPR)*, IEEE Conference, Miami, FL, 1794–1801.

Yang, J., Yu, K., and Huang, T. (2010). "Efficient highly over-complete sparse coding using a mixture model," in *Proceedings of the 11th European Conference on Computer Vision* (Crete: Springer), 113–126.

Zeiler, M., Taylor, G., and Fergus, R. (2011). "Adaptive deconvolutional networks for mid and high level feature learning," in *Proceedings of the IEEE International Conference on Computer Vision* 2011, Barcelona.

Zheng, S., Tu, Z., and Yuille, A. L. (2007). "Detecting object boundaries using low-, mid-, and high-level information," in *IEEE Conference on Computer Vision and Pattern Recognition* (IEEE), Minneapolis, MN, 1–8.

Zhou, X., Yu, K., Zhang, T., and Huang, T. (2010). "Image classification using super-vector coding of local image descriptors," in *Proceedings of the 11th European Conference on Computer Vision: Part V* (Berlin: Springer-Verlag), 141–154.

Zhu, L., Chen, Y., Lu, Y., Lin, C., and Yuille, A. (2008). "Max margin and/or graph learning for parsing the human body," in *2008 IEEE Conference on Computer Vision and Pattern Recognition*, Anchorage, AK, 1–8.

Zoccolan, D., Kouh, M., Poggio, T., and DiCarlo, J. J. (2007). Trade-off between object selectivity and tolerance in monkey inferotemporal cortex. *J. Neurosci.* 27, 12292–12307.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 17 March 2011; accepted: 23 October 2011; published online: 15 November 2011.

Citation: Crouzet SM and Serre T (2011) What are the visual features underlying rapid object recognition? Front. Psychology ***2****:326. doi: 10.3389/fpsyg.2011.00326 This article was submitted to Frontiers in Perception Science, a specialty of Frontiers in Psychology.*